

Weekly Report

Lu Junhua

2017 年 1 月 3 日

Done

- Discussion with Tai-quan Peng. First, about the review of TVCG, resolve some issues. We believe improve writing details will address these problems. Second, about our new data, and potential generalizability problems (单纯说检测外挂, 又将收到reviewer抨击, 说我们太狭隘.)
- Discussion with Yvquan Group.

Besides NN, heuristic method can also be applied to generate features (for detecting abnormal behaviors).

I spotted many flaws in their current method of extracting data. They didn't even look carefully into the sample data. I spent two half-days to filter out all the event types which includes userids. They will re-run their code to extract useful or interesting records.

When to visualize?

To do

- On board of NETEASE Corp.
- Review one course, finish writing course essay.
- Discuss with netease teams, collect their requirements. Discuss one issue with Liu Bai on the TVCG revision. Ask more questions.
- Set about some alternative designs of system. Focus: Feature selection or Events Description.

论文

- CHI 2015 (s|qu)eries: *Visual Regular Expressions for Querying and Exploring Event Sequences* 一个基于正则表达式的可视查询系统. 由于是CHI的文章, 其UI UX设计必须一流. 其主要目的仍然是做查询, 只不过是时间序列上, 用一种类似正则表达式的方式查询. 正则表达式在文中的类比方法是很有趣的, 把event类比成一个alphabet中的字母, 其对应的userid和session id类比于alphabet中行列号; 时间戳类比为文本中一个字母出现的时间顺序. 这样的类比, 通过精心设计的交互, 可以完成许多正则表达式支持的操作.
- 16 *SATORI: A System for Ontology-Guided Visual Exploration of Biomedical Data Repositories*. 本文在知识图谱的基础上进行可视分析. 由于知识图谱略复杂, 在小明的指导下大致了解其内容. 其设

计以及交互方法比较有趣, 不同层次的节点展开、回溯, 作为新的root以后具体查看内部细节、数据分布等等, 可能与计算机文件系统又有那么几分的相似. 如果我们的特征选择也能构成这样的层次特征, 每个特征通过某些手段编码一些分布状况, 视图联动, 是不是也能有一定价值? 如果加上时间, 是不是又是一个创新点?

- KDD 2016 *Catch Me If You Can: Detecting Pickpocket Suspects from Large-Scale Transit Records*
根据公交记录, 结合微博上的小偷公示, 进行挖掘. 虽然要说巨大的新颖也不是在挖掘方法上, 但是在特征选择过程的描述, 以及一个二步框架的描述上就比较厉害, 可能是我们要学习的. 探索时, 用到了1 时间频率直方图、站数直方图2短途旅行3 功能区之间的转移序列(小偷不太正常) 4常去的地方5换乘时间间隔trip时间间隔6历史行为,只用7天数据(这一点很有意思, 之所以这么做是因为要规避非常规线路. 这里就有一个非常规线路, 和小偷(异常)线路的区别, 其实这两种都是异常, 只是目标不同). 文章提出的二步框架, 第一步就是先过滤正常, 再在剩下的人过滤条件异常和小偷. 作者称这样能有些减少false positive.